

QoS-Aware Multi-Armed Bandits

Lenz Belzner
Institute for Informatics
LMU Munich

Thomas Gabor
Institute for Informatics
LMU Munich

Abstract—Motivated by runtime verification of QoS requirements in self-adaptive and self-organizing systems that are able to reconfigure their structure and behavior in response to runtime data, we propose a QoS-aware variant of Thompson sampling for multi-armed bandits. It is applicable in settings where QoS satisfaction of an arm has to be ensured with high confidence efficiently, rather than finding the optimal arm while minimizing regret. Preliminary experimental results encourage further research in the field of QoS-aware decision making.

I. INTRODUCTION

We consider the problem of exploration and exploitation under uncertainty and QoS requirements. Imagine a smart factory control system that is able to provide potential reconfigurations in response to events of change, e.g. on failure detection of a particular machine. In quality critical settings, such a situation may result in downtime until QoS requirements have been reestablished. For example, a factory is required to create products with a guaranteed maximum error rate. At the same time, the confidence about this error rate should be built as fast as possible.

One way to enable verification of QoS requirements at runtime is by performing statistical model checking of the system by using a simulation of the system and its application domain [1]. Here, i.i.d. Monte Carlo simulations of system execution are performed until satisfaction or violation of a particular requirement has been proven up to a given confidence bound.

Given a set of potential reconfigurations in a new situation, QoS-aware automated runtime verification pursues two goals.

- 1) Identify a configuration satisfying QoS requirements.
- 2) Maximize the confidence about this configuration.

This problem yields an exploration vs. exploitation tradeoff: Once a promising configuration has been identified, confidence about its quality should be maximized. However, the system is also interested in configurations with higher quality than the current promising candidate (because QoS confidence can be established faster for these configurations).

Multi-armed bandits (MAB) provide a well-studied formal framework for studying exploration vs. exploitation in decision making [2]. In this paper, we will outline an approach to QoS-aware decision making in the MAB framework. In Section II, we will formally describe the MAB framework and Thompson

sampling, a baseline for MAB decision making based on Bayesian statistics. In Section III, we will describe how to perform QoS-aware Thompson sampling.

II. MULTI ARMED BANDITS

A multi-armed bandit is a set of distributions (e.g. of quality, payoff, utility, etc.). For simplicity, we restrict ourselves to Bernoulli distributions, which return a value of one with a probability of p , and zero otherwise.

A typical task is to identify the optimal arm while maximizing its payoff at the same time. In the case of Bernoulli bandits, the optimal arm i is the one with maximal p_i . A state-of-the-art baseline approach to the bandit problem is Thompson sampling [3], [4]. It builds a distribution about possible values p_i for each arm i , representing the decision makers uncertainty (or beliefs) about the distribution parameter based on its observations.

For Bernoulli bandits, a convenient choice for modeling parameter uncertainty is the Beta distribution with parameters α and β [5]. It is the conjugate prior of the Bernoulli distribution, allowing for efficient posterior computation and analysis [5]. Given an arm i with s_i successes (i.e. s_i times reward one was observed) and f_i failures, and assuming a uniform distribution as prior, the posterior distribution about p_i is given by $Beta(s_i + 1, f_i + 1)$.

Thompson sampling is outlined in Algorithm 1. First, potential values for p_i of each arm are sampled from the current belief distributions. Then, the arm with the best sample is played, and its observed outcomes are updated, effectively changing the belief about its parameter.

Algorithm 1 Thompson sampling for Bernoulli bandits.

- 1: **procedure** THOMPSON SAMPLING
 - 2: Sample \hat{p}_i from $Beta(s_i + 1, f_i + 1)$ for each arm i
 - 3: Play arm i with $\max \hat{p}_i$
 - 4: Update s_i or f_i according to result
-

Despite its simplicity, Thompson sampling has recently attracted research interest due to its theoretical properties and empirical success, showing comparable performance to other state-of-the-art bandit approaches such as UCB [3], [4].

In the context of QoS assessment, each configuration would be represented by an arm of the bandit. The arm's probability of success is the probability that a simulation run of the given

configuration satisfies the QoS requirements. Thompson sampling provides a strategy to identify the optimal configuration wrt. QoS.

However, in situations where it is not necessary to identify the optimal configuration, but rather a configuration that satisfies some QoS requirement with high confidence, standard Thompson sampling tends to put too much effort into optimization, and misses to build confidence in already promising candidate arms. We will outline a solution approach to this problem in the following.

III. QoS-AWARE THOMPSON SAMPLING

A basic form of QoS-aware Thompson sampling (QATS) can be realized by determining the probabilities of QoS violation and satisfaction from the arms' belief distributions. In fact, we are interested in the probability $p_v = P(X \leq q)$ of the true parameter violating the QoS requirement $q \in [0, 1]$. This property can easily be determined from the cumulative density function of the belief distribution.

The probability $p_u = P(X > \hat{p}_i) = 1 - P(X \leq \hat{p}_i)$ that a sampled probability \hat{p}_i from a belief distribution is underestimating the true parameter of an arm is also computable from the belief distribution's cumulative density function.

To solve the exploration vs. exploitation dilemma in a QoS-aware manner, QATS maximizes the odds of underestimation vs. QoS violation. In fact, we prefer large probabilities of underestimation (meaning our belief sample is defensive) while at the same time preferring arms that expose a low probability of QoS requirement violation.

$$o = \frac{p_u}{p_v} \quad (1)$$

Algorithm 2 QoS-aware Thompson sampling.

- 1: **procedure** QoS-AWARE THOMPSON SAMPLING
 - 2: Sample \hat{p}_i from $Beta(s_i + 1, f_i + 1)$ for each arm i
 - 3: Play arm with $\max o_i$ wrt. \hat{p}_i and QoS requirement q
 - 4: Update s_i or f_i according to result
-

QATS is shown in Algorithm 2. We tentatively compared QATS to classic Thompson sampling (TS) in synthetic experiments with promising preliminary results. As an example, consider a four-armed bandit with $p_i \in [0, 0.2]$, instantiated randomly uniform each run. The QoS requirement was set to $q = 0.1$. We evaluated the performance of QATS and TS for 1000 decisions. Figure 1 (top) shows the system's average confidence about QoS satisfaction (i.e. $1 - p_v$) of chosen arms. QATS (blue line) is more confident about QoS satisfaction of chosen arms than TS (orange line). We also measured the cumulative probability of choosing an arm violating the QoS requirement. QATS shows less risk to choose QoS-violating arms. Corresponding results are shown in Figure 1 (bottom).

IV. CONCLUSION

Motivated by runtime verification of QoS requirements in self-adaptive and self-organizing systems that are able to

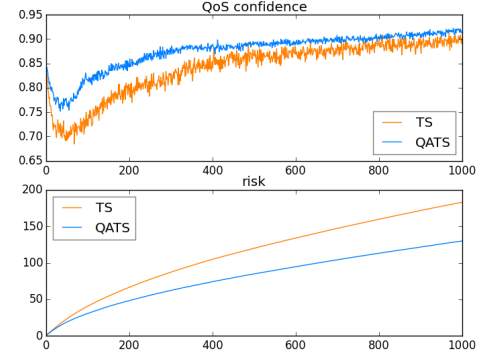


Fig. 1. QATS vs. TS: QoS confidence in sampled arm (top) and cumulative risk of sampling an arm violating QoS requirements (bottom). 350 runs.

reconfigure their structure and behavior in response to runtime data, we proposed QoS-aware Thompson sampling (QATS) for multi-armed bandits. QATS is applicable in settings where QoS satisfaction of an arm has to be ensured with high confidence efficiently, rather than finding the optimal arm while minimizing regret.

Preliminary experimental results are promising and encourage further research in the field of QoS-aware decision making. It would be interesting to investigate theoretical properties of QoS-aware decision making algorithms. Another direction would be to integrate risk measures (as in financial decision making) into Thompson sampling. See [6] for a similar approach based on frequentist confidence bounds. Also, QoS-aware decision making could prove useful in sequential decision making, where one decision changes optimality/quality of subsequent decisions. See [7] for an application of Thompson sampling in Monte Carlo Tree Search. Integration of QoS-awareness into the optimization procedure itself (e.g. the procedure that produces potential system reconfigurations) could allow for even more efficient QoS-aware decision making.

ACKNOWLEDGMENT

The authors would like to thank Matthias Hözl and Martin Wirsing for insightful discussions.

REFERENCES

- [1] I. Lee, O. Sokolsky, J. Regehr *et al.*, "Statistical runtime checking of probabilistic properties," in *International Workshop on Runtime Verification*. Springer, 2007, pp. 164–175.
- [2] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and non-stochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.
- [3] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Advances in neural information processing systems*, 2011, pp. 2249–2257.
- [4] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *COLT*, 2012, pp. 39.1–39.26.
- [5] S. J. Press and J. S. Press, *Bayesian statistics: principles, models, and applications*. Wiley New York, 1989.
- [6] N. Galichet, M. Sebag, and O. Teytaud, "Exploration vs exploitation vs safety: Risk-aware multi-armed bandits," in *ACML*, 2013, pp. 245–260.
- [7] A. Bai, F. Wu, and X. Chen, "Bayesian mixture modelling and inference based thompson sampling in monte-carlo tree search," in *Advances in Neural Information Processing Systems*, 2013, pp. 1646–1654.